COMP-767: Reinforcement Learning - Assignment 1

Posted Thursday, January 17, 2019
Due Tuesday, January 29, 2019

The assignment can be carried out individually or in teams of two. You have choices on both parts of the assignments.

1. **Bandit algorithms [50 points]**

   Choose **one** of the following topics.

   (a) Best-arm identification: also sometimes called *pure exploration*, this is a problem formulation in which the

   For this part of the assignment you will have to read the following paper:

   Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting, Kevin Jamieson and Robert Nowak, CISS, 2014:

   https://people.eecs.berkeley.edu/ kjamieson/resources/bestArmSurvey.pdf

   The paper describes two different classes of algorithms for this problem. Your task is to: (a) summarize the main results in the paper; (b) Reproduce the results in Figure 1 (c) Perform the same empirical comparison on the bandit problem provided in the Sutton & Barto book (which we discussed in class). Do not forget to average your results over multiple independent runs. (d) discuss in a short paragraph a concrete application in which you think regret optimization would be more useful than best arm identification.

   (b) We discussed in class Thompson sampling, which is a very useful Bayesian algorithm for bandits. We mentioned briefly regret results. In this part of the assignment, you should read the following paper by Agarwal and Goyal (2012), which provides early prior-independent regret bounds for Thompson sampling:

   https://arxiv.org/pdf/1209.3353v1.pdf

   You need to write a short summary (max 3 pages in latex format of your choice) of the result and the main steps and ideas in the proof. Feel free to add some background if you think it would be necessary to understand their approach.

2. **Markov Decision Processes and dynamic programming [50 points]**

   Choose **one** of the following topics.

   (a) Implement and compare empirically the performance of value iteration, policy iteration and modified policy iteration. Modified policy iteration is a simple variant of policy iteration in which the evaluation step is only partial (that is, you will make a finite number of backups to the value function before an improvement step). You can consult the Puterman (1994) textbook for more information. You should implement your code in matrix form. Provide your code as well as a summary of the results, which shows, as a function of the number

of updates performed, the true value of the *greedy policy* computed by your algorithm, from the bottom left state and the bottom right state. That is, you should take the greedy policy currently considered by your algorithm and compute its exact value.

To test your algorithm, use a grid world in which, at each time step, your actions move in the desired direction w.p. p and in a random direction w.p. (1-p). The grid is empty, of size $n \times n$. There is a positive reward of +10 in the upper right corner and a positive reward of +1 in the upper left corner. All other rewards are 0.

You need to test your algorithm with two different values of p (0.9 and 0.7) and with two different sizes of gird ($n = 5$ and $n = 50$). Explain what you see in these results.

(b) We mentioned briefly in class the linear programming approach to solving MDPs. One reason why this approach is interesting is that it can work with either a primal or a dual formulation. For this part of the assignment, you will read a paper by Wang et al, which attempts to leverage these ideas but in dynamic programming:

https://era.library.ualberta.ca/items/6ef4eab6-acb1-48f3-a188-8c6f2aa39c2d/view/7145a384-5ea5-4d7e-abe0-d29e013b8921/TR07-10.pdf

You need to write a short summary (max 3 pages in latex format of your choice) of the intuition of their approach, what is their "primal" and "dual" formulation, and what are the main theoretical insights of the work. Feel free to add some background if you think it would be necessary to understand their approach. In one paragraph, please propose one improvement to this work.